PECAN
A fast, novel 16S rRNA gene sequence non-clustering based taxonomic assignment tool

Johanna B. Holm, Pawel Gajer, Jacques Ravel
Institute for Genome Sciences, University of Maryland Baltimore School of Medicine

PECAN
Per sEquenCe tAxonomic assigNer

Background

Clustering of sequences into Operational Taxonomic Units (OTUs) has become a mainstream approach to facilitate taxonomic classification of large numbers of 16S rRNA gene sequences. This is partly due to the high computational requirements for processing each sequence in increasingly large datasets. A primary focus of the field has been development and improvement of OTU-based sequence clustering methods that rely on distances between each pair of sequences in a dataset. Following OTU-based clustering, representative sequences are commonly classified using tools such as the RDP Naïve Bayesian Classifier (Wang *et al.* 2007), and the resulting classification transitively assigned to all sequences comprising that OTU. However, problems with this strategy exist (Nguyen *et al.*, 2016). Here, we present **PECAN**, a novel per sequence taxonomic assigner which quickly and accurately classifies millions of 16S rRNA gene sequences using higher order Markov Chain models built from a user-specified set of reference sequences.

Methods & Usage

1. Obtain, align, & truncate reference sequences to variable region(s) of choice

- RDP database, MAFFT v 7.222, & mothur v 1.36.0

Reference sequence curation is an ongoing project. We seek to correct sequence annotation AND ensure the loss of information due to sequence truncation is reported. Help out!

2. Produce phylogenetic tree

- FastTree JC + CAT

3. Curate taxonomic annotation curation (Figure 1)

- An internal pipeline to automatically assign corrected taxonomic annotations.

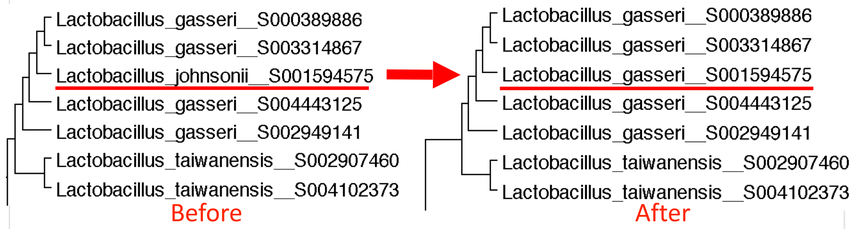

Figure 1. Example of taxonomic annotation curation. Prior to curation, sequence S001594575 was annotated as *L. johnsonii.* It is corrected to *L. gasseri* after alignment-dependent curation.

4. Curation produces a taxonomic tree containing all sequences (Figure 2). For each species, genus, etc., construct 7th order Markov Chain model (Table 1).

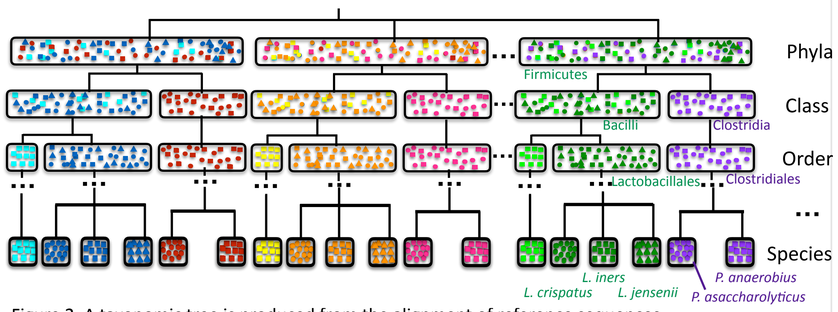

Figure 2. A taxonomic tree is produced from the alignment of reference sequences

Table 1. Markov Chain models of the probability of observing a particular base following a particular 7mer using *Lactobacillus* spp. as an example.

	A | 7mer$_a$	C | 7mer$_a$	G | 7mer$_a$	T | 7mer$_a$	A | 7mer$_b$	C | 7mer$_b$
L. iners	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
L. jensenii	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
L. crispatus	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
g_*Lactobacillus*	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	...	...	...	...	...	...
f_Lactobacillaceae	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	...	...	...	...	...	...
o_Lactobacillales	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	...	...	...	...	...	...
c_Bacilli	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	...	...	...	...	...	...
p_Firmicutes	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)

"For *L. iners*, this is the probability of observing a C given the previous 7 nucleotides" (here, labeled as "7mer$_b$")

5. The reference taxonomic tree is represented by Markov Chain models (Figure 3).

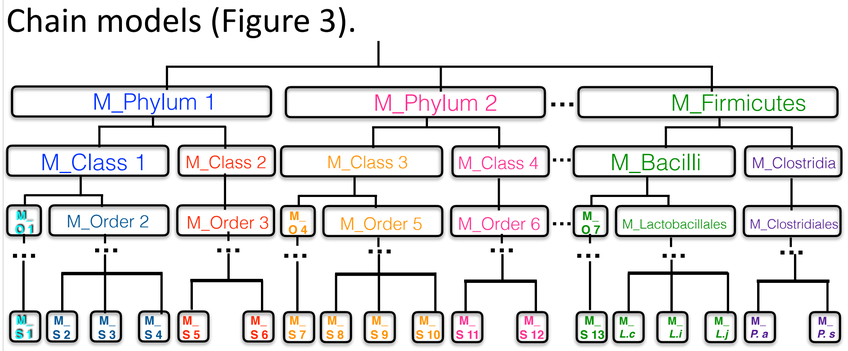

Figure 3. Schematic illustrating the taxonomic tree built from curated reference sequence information is represented by Markov Chain models annotated as "M_".

6. Confidence in taxonomic assignments comes from error models generated during reference model construction.

- For each reference node, 1,000 sequences are randomly generated from that node's Markov chain model M.
- Posterior probabilities of M and its sibling nodes are computed.
- The distributions of the resulting posterior probabilities are compared (Figure 5).
- A threshold of posterior probability that minimizes misclassification error is computed (Bayes decision theory).

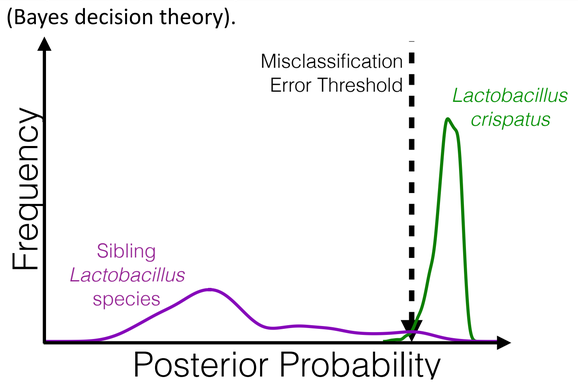

Figure 4. Error models are used to ensure correct assignments by PECAN. Here, *Lactobacillus crispatus* is used as an example. The distributions of the posterior probabilities of the *L. crispatus* model (green) and the siblings of *L. crispatus* model (purple) are computed. The taxonomic assignment misclassification error threshold is the intersection of these two distributions.

7. Query sequence taxonomic assignment with PECAN

- First, PECAN classifies at the phylum level by calculating the posterior probability of the sequence using each phylum's Markov chain model (Figure 4).
- **The sequence is assigned to the phylum whose model produces the highest posterior probability.**
- The assignment is accepted if this posterior probability is above the assignment error threshold (Step 6, Figure 4).
- This process is repeated at each taxonomic level until either the sequence is classified to a species, or the threshold requirement is not satisfied.

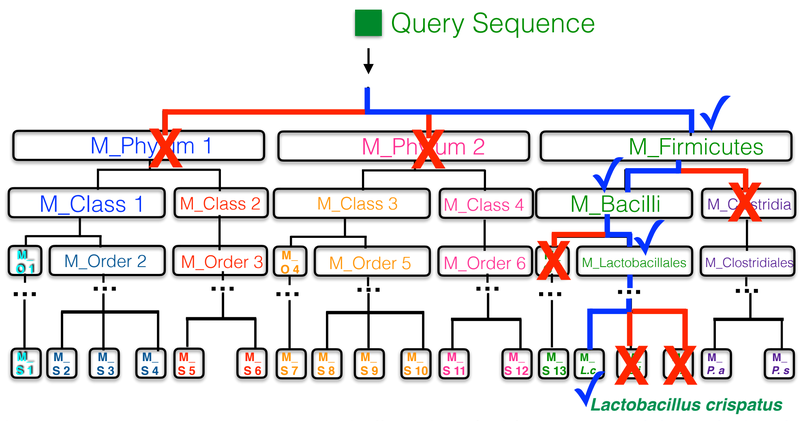

Figure 5. A query sequence is taxonomically assigned with PECAN. Models producing the highest posterior probability and satisfying the minimum error threshold requirement are accepted (blue check marks). The remaining models are rejected (red X's). Only nodes within the higher level assignment are considered. In this example, the query sequence is classified as *L. crispatus*.

8. Accuracy of PECAN was estimated using 10-fold cross validation

- We tested PECAN on our curated reference database containing 1727 unique sequences representing 128 bacterial taxa from the vaginal environment.
- 10-fold cross validation testing was performed and we report median accuracy ± median absolute deviation (MAD) of classification.

9. The speed of PECAN assignment was esimated with **15.3 million** amplicon sequences (16S ribosomal RNA gene variable regions 3 & 4, ca. 450 bp) from 383 vaginal samples. Amplicons were generated using a 2-Step PCR amplification protocol and sequenced on an Illumina HiSeq 2500.

Results

PECAN Accuracy:

Median Accuracy ± MAD: **97.31**% ± 0.81

PECAN Speed:

15,377,006 sequences classified in 14.93 min on an Intel(R) Xeon(R) CPU 5160 at 3.00GHz

> 1 million sequences classified per minute

SeqID	Taxonomic Assignment
VM399_0	Gardnerella_vaginalis
VM155_1	Candidate_Division_TM7_vaginal
VM192_2	Leptotrichia_amnionii
VM119_3	Lactobacillus_crispatus
VM197_4	Lactobacillus_crispatus
VM120_5	Lactobacillus_crispatus
VM398_6	Lactobacillus_crispatus
VM132_7	Gardnerella_vaginalis
VM170_8	Lactobacillus_jensenii
VM197_9	Lactobacillus_jensenii
VM119_10	Lactobacillus_crispatus
VM171_11	Lactobacillus_iners
VM147_12	Prevotella_genogroup_1
VM130_13	Lactobacillus_jensenii
VM116_14	Lactobacillus_iners
VM130_15	Lactobacillus_jensenii
VM189_16	Lactobacillus_iners
VM115_17	Lactobacillus_crispatus
VM176_18	Lactobacillus_iners
VM160_19	Gardnerella_vaginalis
VM127_20	Lactobacillus_crispatus
VM399_21	Parvimonas_micra
VM144_22	Sneathia_sanguinegens
VM398_23	Finegoldia_magna

Figure 6. Example PECAN output file

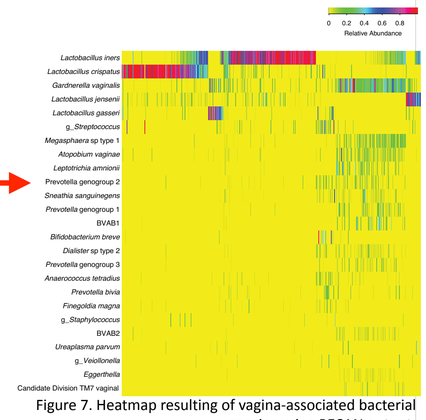

Figure 7. Heatmap resulting of vagina-associated bacterial species using PECAN output.

Summary

- PECAN classifies over 1 million amplicon sequences per minute with high accuracy
  - ✳ Built with C/C++ and uses a taxonomic tree to reduce required comparisons between query sequences and reference models
  - ✳ Taxonomic assignment speed is a linear function of query sequence length
- PECAN removes the necessity and issues associated with OTU-based sequence clustering
- PECAN is applicable to any microbiome dataset by customization and curation of the reference sequences by variable region and/or environment
- Accuracy of a classifier critically depends on the accuracy of the taxonomy of reference sequences. The truncation of 16S rRNA gene sequences results in the loss of information. This loss of information should be explicitly measured and reported. This has guided the development of our sequence curation pipeline.

Learn & Use PECAN

PECAN
Per sEquenCe tAxonomic assigNer

View a digital version of this poster AND get PECAN @
http://ravel-lab.org/pecan

Funding

The development of PECAN was supported by the National Institute of Allergy and Infectious Diseases and the National Institute of General Medical Sciences of the National Institutes of Health under awards numbers U19AI084044, R01AI116799 and R01GM103604.

References
Nguyen, NP, Warnow, T, Pop, M, White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npg Biofilms & Microbiomes* 2: 1-8.
Wang, Q, Garrity, G, Tiedje, J, Cole, J (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA sequences into the new bacterial taxonomy. *AEM* 73(16): 5261-5267.
Yamada, KD, Tomii K, Katoh, K, (2016). Application of the MAFFT sequence alignment program to large data - reexamination of the usefulness of chained guide trees. *Bioinformatics* 2016: 1-6.
Katoh K, Misawa K, Kuma K, Miyata, T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acids Res.* 30(14): 3059-3066.
Schloss, PD, Westcott, SL, Ryabin *et al.* (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *AEM* 75 (23): 7537-7541.
Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution 26:1641-1650
Illumina protocol: http://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf